General comments:

The authors have developed a coupled data assimilation (DA) system, which consists of an earth system model (ESM) and parallel DA framework, and have conducted DA experiments using the ocean component. Although the authors might make large efforts to construct this system, this paper includes many problems as seen in specific comments. At least two critical points indicated below should be improved.

First, the authors have indicated that the advantage of this paper is the scalability easy to choose combinations of multiple components in coupled DA systems. However, the authors have performed the DA experiments with the ocean component only, and have not demonstrated the advantage of scalability. To improve this point, the authors need to conduct DA experiments using various coupled systems.

Second, the authors have not sufficiently described the experimental setting and validation methods, to provide an accurate understanding for the readers. Especially in the experimental setting, there are serious problems with the ensemble size and how to generate observations in twin experiments. For the former, the ensemble sizes of 8 and 16 in this study are too small for high-dimensional systems, and for the latter, the no-bias assumption between the forecasts and observations is not satisfied.

Long periods would be necessary for the authors to re-construct the experimental setting and perform the experiments, and therefore I suggest between “major revision” and “reject” with approval of re-submission.

Specific comments:

#1: The novelty of this paper is to establish the new coupled DA system easy to include multiple components such as atmosphere, ocean, sea-ice, and biogeochemistry. However, the authors performed experiments using only the ocean DA system, with the description of “extension to coupled ESMs (e.g., ocean-atmosphere-biology etc) is trivial” in L56 in Section 1. Consequently, the authors have not demonstrated the scalability that this new DA system can be easily extended to the coupled DA systems. At least, experiments using various coupled DA systems are necessary to prove it.

#2: The descriptions for the experimental setting are not polished well. For example, from “using a simple idealized test with manufactured ‘observations’ in L131 in Section 3, we could understand that the authors have performed twin experiments, but this description is not straightforward. Although the authors described that outputs from the free run are used for generating observations, observations in twin experiments should be generated by adding random noises to the true values from the nature run. Please use suitable

terminology (twin experiment, nature run, free run, etc.). Therefore, subsection 3.1.2 should be modified overall.

#3: Wind forcing has impacts on the ocean through wind stress and turbulent heat fluxes, but the authors apply “the surface wind that mixes the water” to the ocean model as described in L140-141in subsection 3.1.1. It is not clear how the authors apply the wind forcing.

#4: It seems that the forecast ensemble for initial conditions is generated based on Pham et al. (2001). Is this procedure applied to each assimilation cycle? If not, the filter divergence is likely to occur because the ocean is mainly controlled by the atmospheric forcing and the ensemble spread is under-dispersive. To avoid filter divergence, the perturbed atmospheric and lateral boundary conditions are generally applied in ocean data assimilation systems (e.g., Penny et al. 2013). To confirm that the filter divergence does not occur in this system, the authors should draw figures showing ensemble spread.

#5: It appears that the authors generate the initial forecast ensemble mean by adding- -1 °C to the nature run in the twin experiments. This indicates that there are biases between the forecasts and observations. Since no biases are assumed in the formulation of the Kalman filter, this experimental setting is not appropriate for EnKF.

#6: As described in the last paragraph in subsection 3.1.2, the authors use three kinds of observation errors. In twin experiments, prescribed observation errors of 0.15 °C are generally used. Therefore, the twin experimental settings using the different observation errors are not reasonable. Furthermore, uniform observation errors in the first option would indicate that the same observation errors are used for different observations with different units. This does not make sense.
It is not clear to describe the observation errors for the twin or real experiments, and the authors should summarize the experimental settings for twin and real experiments into different subsections.

#7: The ensemble sizes of 8 and 16 in the twin and real experiments, respectively, are too small for high-dimensional ocean DA systems. Even in the Lorenz96-LETKF system with 40 variables, the low limit of the ensemble size is 8–10. To suppress the pseudo correlation, more than several tens of ensemble members, preferably more than 100 would be necessary for this system.

#8: The authors described that the EnKF with the localization has better accuracy than that without the localization in subsection 3.1.4. As shown in Kondo and Miyoshi (2016), this is not true for all cases. If the ensemble size is so large, the EnKF without localization can surpass that with localization.

#9: MAEs in the experiments using EnKF with and without the localization are about 0.2 and 0.6 °C and largely different. This is inconsistent with the description of "Both global and local filters achieved similar results." in L184 in subsection 3.1.4. To clarify the accuracy differences between EnKF with and without the localization, the statistical test should be applied.

#10: The descriptions of the role of localization in subsection 3.1.4 are incorrect. Although the authors describe that localization limits observation impacts and results in better accuracy, localization is important to suppress the pseudo correlation as described in the 6th major comment.

#11: The authors have not described what kinds of inflation (additive/multiplicative, RTPP, RTPS) is adopted in experiments with LETKF and LESTKF, and have not conducted sensitivity experiments to investigate the impacts of the inflation methods on the accuracy. However, the authors described that the LESTKF has an advantage for ensemble spread compared with LETKF. Furthermore, the authors have not specified the computational cost but described that the computational cost is slightly better in the LESTKF than the LETKF. Could you give more evidence for the ensemble spread and computational costs?

#12: Because of the large differences between the in-situ buoy and satellite-based analysis SSTs, only the satellite-based products are assimilated in this study. However, on the assumption that in-situ observations have smaller errors, satellite SSTs estimated from the infrared and microwave radiance are validated relative to in-situ observations. This is the general procedure to create satellite-based observational datasets. Furthermore, satellites cannot observe cloudy and rain areas by infrared and microwave sensors, respectively. Therefore, errors in satellite SSTs would be larger. The authors should show clear evidence that the assimilated SST analysis products have smaller errors than the in-situ buoys. My suggestion is to use independent buoy data for validation.
It is better to use satellite observations themselves rather than analysis products because

the analysis errors might be correlated with the forecast errors. The Kalman filter is derived from the assumption that observation errors do not correlate with the forecast errors.

#13: Throughout the subsection 3.2.2, the authors have not described what data are used for the validation, and therefore I could not accept the results. To clarify this point, it might be better to make subsection to describe the validation method.
For the real DA experiments, the authors use “Mean Absolute Errors (MAEs)” for validation. Can the authors estimate errors relative to true values? If not, the MAEs are not appropriate. Mean absolute deviations (MADs) rather than MAEs can be estimated in real DA experiments.

#14: The authors described “the Kuroshio induced upwelling near northeast corner of Taiwan” in L257 in subsection 3.2.2. Could you investigate the mechanisms of how the Kuroshio results in upwelling?
The authors might intend SST cooling caused by the upwelling related to the typhoon passage. However, the SST cooling can be caused by strong turbulent heat releases as well. Consequently, the mixed layer heat budget analysis is necessary to demonstrate that the SST cooling results from the upwelling.

#15: The authors described “Overall, the two cases have similar SST skills.” in L257-258 in subsection 3.2.2. However, as shown in Fig. 10, Case A has better accuracy than Case B, and this is inconsistent with the description. This is the same as the descriptions about the results of SLA.

#16: The Kuroshio undergoes seasonal variations with larger (smaller) transport in boreal summer (winter) as well as interannual variations. Although the authors compared the analytical Kuroshio transport in June–July 2016 with the observational-based climatology reported by the previous studies, the experiment periods are short, and therefore it is not appropriate to compare with the climatology.
The authors described that the analytical warm SSTs increase the Kuroshio transport by increasing the density vertical gradient. However, the thermal wind equation shows that the horizontal temperature gradient is related to the vertical wind shear. Therefore, the authors’ explanation does not make sense, and the authors should give evidence to show the mechanism.

Technical corrections:

I have not indicated all of points to be corrected, but I list some of them. I ask the authors to describe more carefully to facilitate the review process.

L31: “and is significantly more efficient” should be “, is significantly more efficient”
L31: Could you specify more about “the nonlinear feedback”?
L33: “nonlinear” might not be necessary in “nonlinear particle filters”.
L38: Referring “Fig. 2” at first is not reasonable.
L56 and others: “etc” should be “etc.”.
L68: Replace “in an obvious way” with the end of the sentence.
L69 and others: Countable and uncountable nouns of “structure” have different meaning. The authors intend to uncountable noun of “structure”, and “structures” is not appropriate.
L82: Insert full spell of “SCHISM”.
L82 and L104: Change the abbreviated to unabbreviated URL.
L88: “in observation data … for each type” might be “observations … for each observation type”.
L89: “map” might be “project”.
L90-91: Generally, observations within cut off scale are used for assimilation. Here, the authors only use observations within localization scale. The authors might mix the definition between localization and cutoff scales.
L94: Please specify what variables are outputted in output_netcdf_pdaf (ex. analysis ensemble mean and spread).
L94: Why are only the analysis ensemble mean outputted? The ensemble spread is also important to monitor the filter divergence.
L96: Since “significant” is used for only the results with statistical test, “Significantly” is not appropriate.
L98: “model forcing” should be “external forcing”.
L100: The grammar of “At the lowest level … model component” is not correct.
L103: Spell out “LSC”.
L 164: Is the localization scale of 500 m in the horizontal and/or vertical localization scale ?
L165 and others: Abbreviation of “Fig.” at the start of sentences should be spell out.
L166: “the difference between before and after assimilation” should be “the differences between forecast and analysis ensemble mean (i.e., analysis increments)”. Remove “the ‘the forecast’ is … after DA’.

L210: Perhaps HYCOM provides outputs from the model or analyses. Since “data” are used as “observations”, “data” should be removed.
L213: Compared to what, does the foreward model have larger errors ?
L228: Argo data are used in this system, but are the temperatures only assimilated? Why are the salinity observations assimilated?
L230: “observation error” consists of measurement and representation errors. Do satellite datasets provide both errors?
L231 and others: “PSU” is a practical unit and not used in scientific papers.
L234: Do “specified influence range” mean the cutoff scale?
L255: What does “deeper region” indicate?
L257: Insert “that” between “well” and “the Kuroshio”.
L338: Data assimilation cannot improve “the model skill”.

Table 1: Add Bishop et al. (2001) to the reference of ETKF.
Figure 5: Observation points cannot be found in the left upper panel.
Figure 7: Better to show the observation frequency or total observations by Argo profiling floats rather than the snapshots.
Figures 6, 10, 12, and 14: Use date rather than counted days from the reference date.
Figures 7, 8, 9, 11, and 13: Use longitude, latitude, depth for the axis and labels rather than x-, y-, z-axis, respectively.

References:
Penny SG, Kalnay E, Carton JA, et al (2013) The local ensemble transform Kalman filter and the running-in-place algorithm applied to a global ocean general circulation model. Nonlinear Process Geophys 20:1031–1046. https://doi.org/10.5194/npg-20-1031-2013

Kondo K, Miyoshi T (2016) Impact of removing covariance localization in an ensemble Kalman Filter: Experiments with 10 240 members using an intermediate AGCM. Mon Weather Rev 144:4849–4865. https://doi.org/10.1175/MWR-D-15-0388.1